**Final Project Report**

自硕21 崔晏菲 2021210976

注：因为我不会R语言，所以代码都是用Python写的。代码文件见FinalProject-code.ipynb

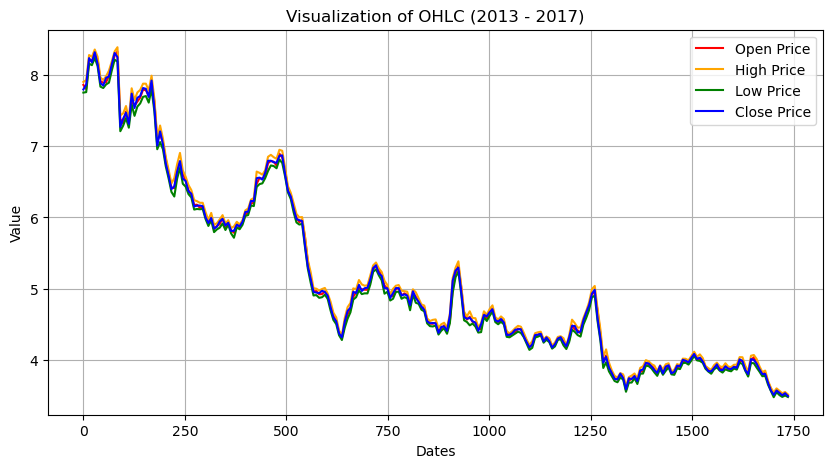
1. **玉米价格建模**
   1. **数据预处理**

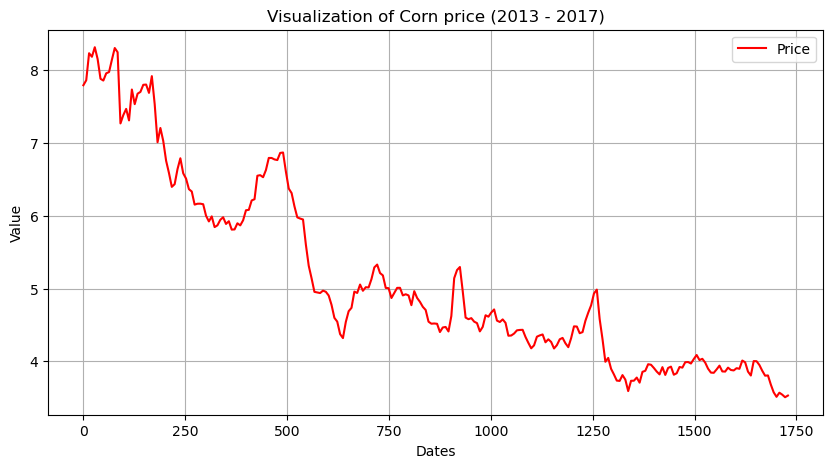
读取三个表格的数据，发现corn\_OHCL2013-2017.txt中的日期比另外两个表格多月了一个2017-10-08，故删去。

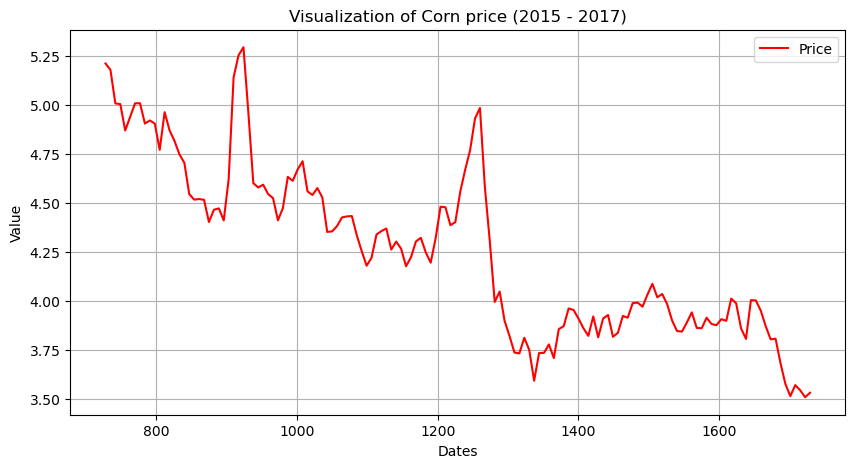
接着，将文字类型的日期转换成int类型的整数，我选择以2013-01-06为第0天，将其他日期转换成与2013-01-06相间隔的天数。

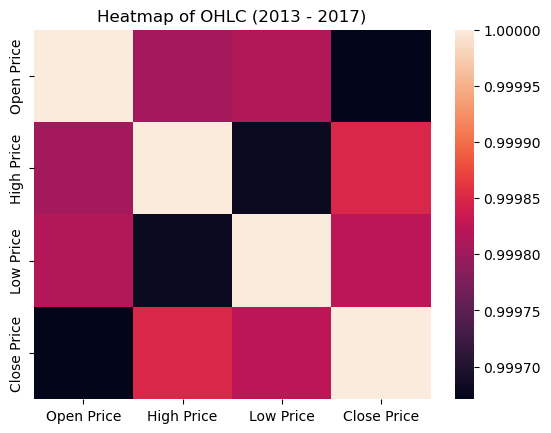
* 1. **数据可视化**

首先是最基本的折线图展示：



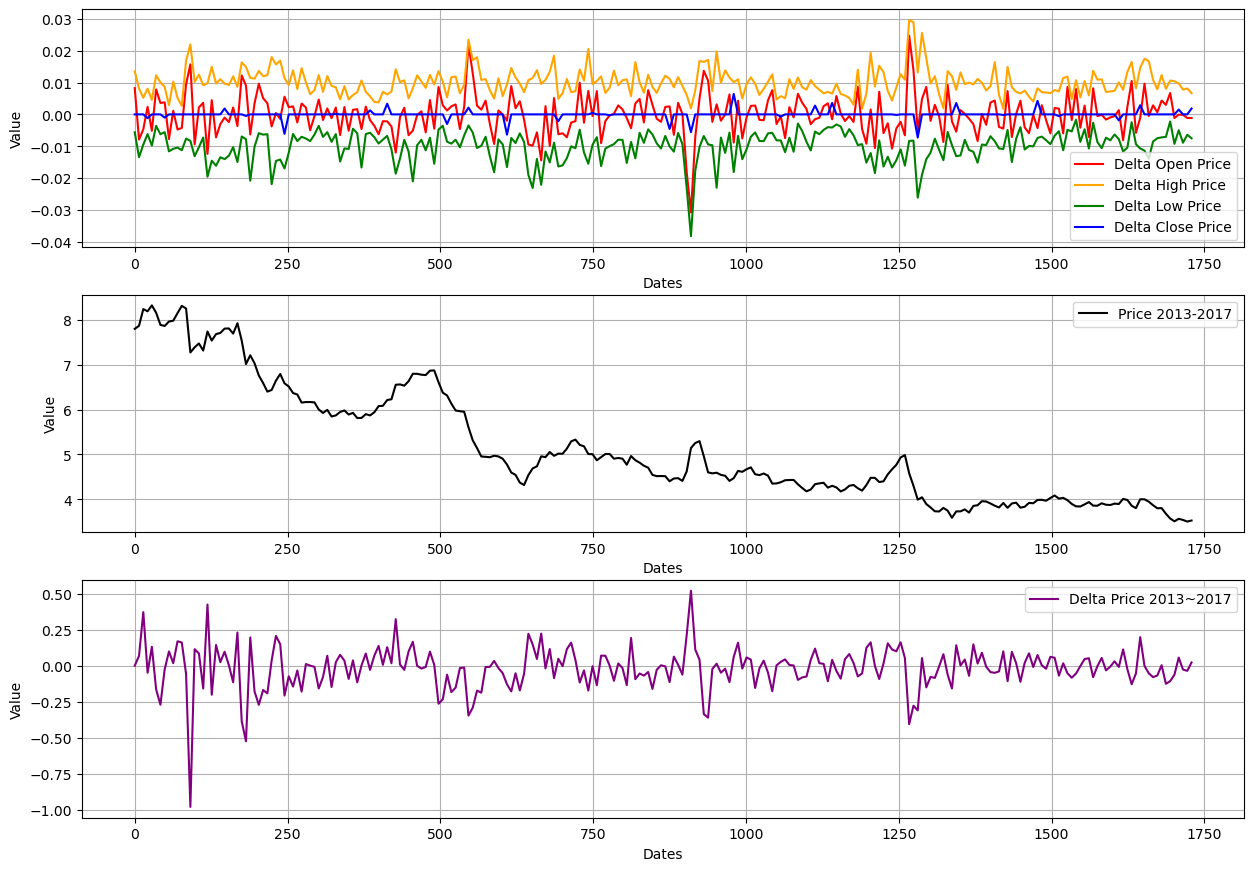




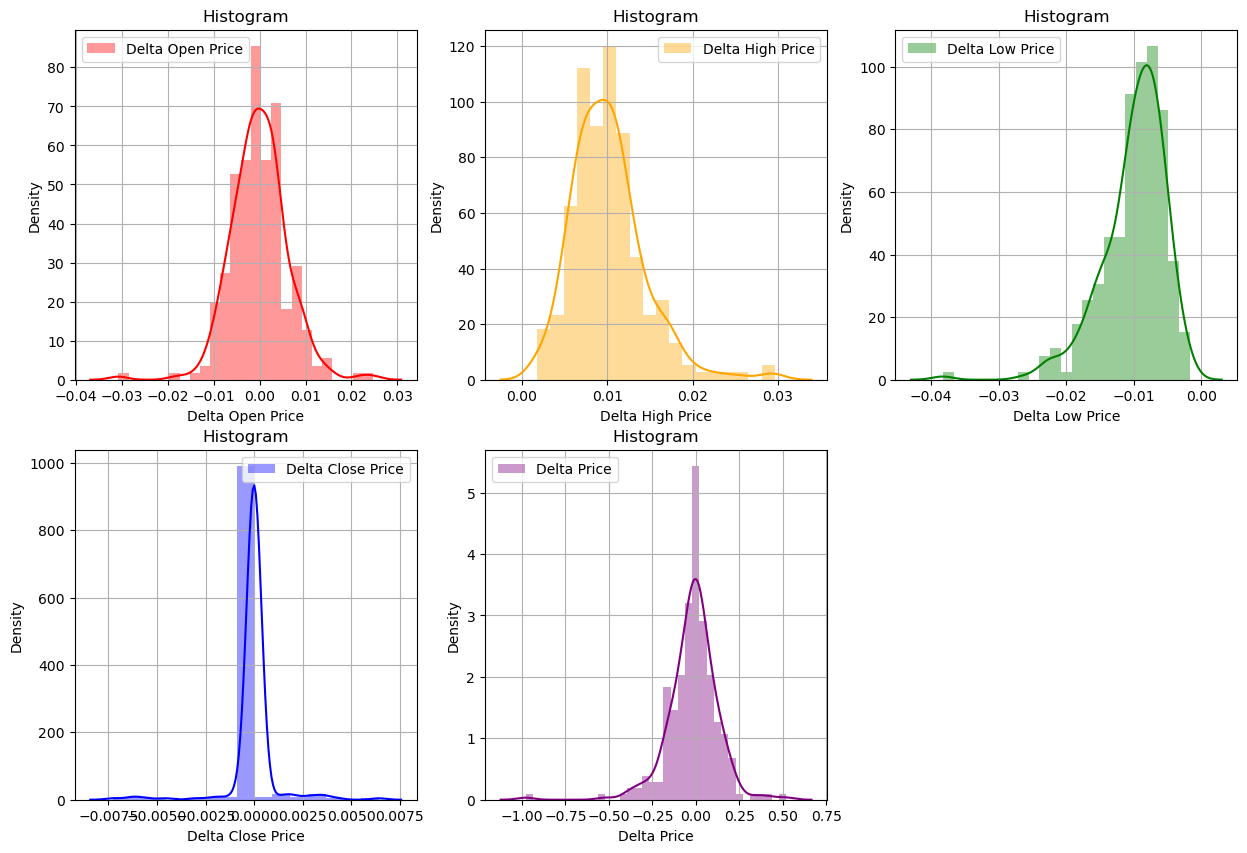


以上是三个表格各自的折线图，以及corn\_OHCL2013-2017.txt中四列数据的相关性热图。可以看到玉米价格曲线整体上基本可以按照指数函数的方式拟合，但在小范围内波动非常大；以及corn\_OHCL2013-2017.txt中的四个变量相关系数都很高，且取值过于接近。这样直接建模没有什么实际意义，并不能预测玉米价格。

因此，我以corn\_2013-2017.txt中的玉米价格为每日的基准价格，计算每日的开市价格、闭市价格、最高价格、最低价格在基准价格基础上的变化量，即得到新的折线图如下：



上图是一个对照图，其上面部分是玉米每日的开市价格、闭市价格、最高价格、最低价格在基准价格基础上的变化率；中间部分是每日玉米价格的折线图；最下面的部分是玉米价格与上一周相比的变化率，数值大于0时说明这周相比于上周上涨了，小于0说明下降了，这三部分共用同一个横坐标。在图上可以直接的看出来，红色线和蓝色线实在围绕玉米基准价格波动；而当绿色线和橙色线到达局部极值点时，玉米的价格往往也正好处于极值点。

接着查看这些变化率的分布情况，如下图所示：  
  
可见，Delta Open Price、Delta Close Price和Delta Price大概服从正态分布；Delta Low Price和Delta Close Price大概服从卡方分布。

* 1. **问题建模**

到这里，我们总算想到了一个具有实际意义的问题：**假设我是一个投资玉米的人，那我如何判断自己应该在什么情况下买入，什么情况下卖出呢？一个**很自然的想法是，当判断玉米价格正在上涨时买入，当判断玉米价格正在下跌时卖出。那么问题就转换成了，如何判断玉米价格正在上涨还是正在下跌？

由于玉米价格处于上涨状态还是下跌状态是不可观测的，能观测的只有过去的价格数据。这刚好可以用隐马尔科夫模型来建模。

记隐状态代表上涨状态，代表下跌状态。记，并且服从正态分布先验，服从卡方分布先验。记，状态转移矩阵为，初始隐状态的概率分布为，则有

记为当（）时的条件概率，则有：

记为观测到前个时刻的数据且（）的概率，则

记为给定第个时刻的的条件下，观察到后续的观测值为的概率，则

则时刻从状态迁入状态（）的概率为：

且序列的概率为:

接下来使用EM算法来进行参数估计与推断。

E-step:

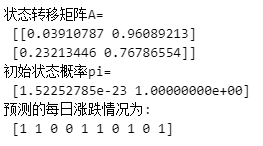
M-step:

* 1. 模型估计

在模型估计的过程中，一开始我直接使用全部的可观测数据进行EM算法，但很快就会发现结果总是溢出。经过排查，代码中不存在bug，EM算法的第一遍迭代还是可以正常运行的。问题其实是出在计算发射概率上面了，因为这里我直接用的概率密度分布函数的乘积，这样会导致的返回值远大于1，接着在前向和后向的动态规划中反复相乘导致数值溢出。这在一条长达248的马氏链中是无法避免的。

因此，隐马尔科夫模型只能处理很短的一小段时间段，并不能一次性处理全部时间段。接下来我全部使用的长度为10天的马氏链来进行模型的训练和预测。

例如，我使用2013年前10条数据进行训练，可以得到的结果是：



但是因为HMM模型在这里用的是无监督学习方法，所以对于HMM来说，0和1是对称的，无法确切指出今日是涨还是跌，只能知道今天的涨跌情况和其他天数是否一样。

* 1. 模型推断

接着我在这248条数据中全都用HMM模型进行了涨跌情况的预测，并与ground truth进行了比较，结果发现，预测得非常不准，完全没有统计学上的意义。

果然，这个实验又一次证明了无数前辈先贤的血泪经验——统计学是没办法预测股价的。